NONVOLATILE MEMORY CELL HAVING FLOATING GATE, CONTROL GATE AND SEPARATE ERASE GATE, AN ARRAY OF SUCH MEMORY CELLS, AND METHOD OF MANUFACTURING

TECHNICAL FIELD

[0001] The present invention relates to a nonvolatile floating gate memory cell having a control gate and a separate erase gate, an array of such cells, and a method of manufacturing.

BACKGROUND OF THE INVENTION

[0002] Nonvolatile memory cells have a floating gate for the storage of charges thereon to control the conduction of current in a channel in a substrate of a semiconductive material is well known in the art. See, for example, U.S. Patent No. 5,029,130 whose disclosure is incorporated herein by reference in its entirety. In U.S. Patent No. 5,029,130, a split gate nonvolatile memory cell having a floating gate with source side injection and poly to poly tunneling is disclosed. The memory cell has a first region and a second region with a channel region therebetween with the channel region having a first portion and a second portion. A floating gate is disposed over a first portion of the channel region is insulated therefrom and controls the conduction of current in the channel region depending upon the charges stored in the floating gate. A word line/erase gate is disposed over a second portion of the channel region and is insulated therefrom and controls the conduction of current in the second portion of the channel region. The cell is programmed when electrons through the mechanism of hot electron channel injection are injected from the channel region onto the floating gate. The cell is erased by electrons from the floating gate tunneling to the erase gate through the mechanism of Fowler-Nordheim tunneling. The floating gate is characterized by having a sharp tip to facilitate the tunneling of electrons from the floating gate to the control gate. In U.S. Patent No. 5,029,130, the control gate/erase gate performs two functions. First, it controls the conduction of current in the second portion of the channel region during the operations of programming and read. Secondly, it is supplied with a high voltage during the erase operation to attract the electrons from the spaced apart and insulated floating gate. These two functions have compromised the design of a single member which must perform both functions. Specifically, during programming and read, the word line/control gate receives low voltage whereas during erase, it must receive a high voltage.

[0003] It is therefore, an object of the present invention to overcome this and other difficulties.

SUMMARY OF THE INVENTION

[0004] Accordingly, in the present invention, a nonvolatile memory cell comprises a substrate of substantially single crystalline semiconductive material having a first conductivity type. A first region of a second conductivity type is in the substrate. A second region of the second conductivity type is in the substrate spaced apart from the first region. A channel region is between the first region and second region with the channel region having a first portion and a second portion. A control gate is insulated from the second portion of the channel region. A floating gate is adjacent to the control gate and is insulated therefrom. The floating gate is also insulated from the first portion of the channel region. The floating gate has a tip which is closest to the control gate. An erase gate is insulated from the control gate and the tip of the floating gate. An insulating material is between the tip and the erase gate to permit charges to tunnel from the tip to the erase gate.

[0005] The present invention also relates to an array of the foregoing described nonvolatile memory cells. Finally, the present invention relates to a method of manufacturing an array of nonvolatile memory cells.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] Figure 1A is a cross-sectional view of a first embodiment of a nonvolatile memory cell of the present invention, in which the nonvolatile memory cell is unidirectional in operation and is formed on a planar surface of a semiconductor substrate. Figure 1B is a schematic circuit diagram of an array of memory cells of the first embodiment shown in Figure 1A.

[0007] Figure 2A is a second embodiment of a non-volatile memory cell of the present invention in which the nonvolatile memory cell is bi-directional in operation and formed in a trench and a planar surface portion of a semiconductor substrate. Figure 2B is a schematic circuit diagram of an array of memory cells of the second embodiment shown in Figure 2A..

[0008] Figures 3A-3J are the steps showing a method of manufacturing the nonvolatile memory cell of the first embodiment shown in Figure 1A and the array shown in Figure 1B. Figure 3K is a cross-sectional view taken along the line 3k-3k in Figure 4.

[0009] Figure 4 is a top view of an array of nonvolatile memory cells of the first embodiment shown in Figure 1A in which cells are offset to accommodate strapping lines.

[0010] Figures 5A-5K are cross-sectional views showing the process of making the nonvolatile memory cells of the second embodiment shown in Figure 2A, and the array shown in Fig. 2B.

DETAILED DESCRIPTION OF THE INVENTION

[0011] Referring to Figure 1A, there is shown a cross-sectional view of a first embodiment of a nonvolatile memory cell 10 of the present invention. Similar to the cell shown and described in U.S. Patent No. 5,029,130, whose disclosure is incorporated herein in its entirety by reference, the memory cell 10 is formed in a substantially single crystalline semiconductor substrate 12, such as silicon. The substrate 12 is of a first conductivity type. The substrate 12 also has a planar surface 8. Within the substrate 12 is a first region 14 of a second conductivity type. A second region 16 of a second conductivity type is spaced apart from the first region 14. Between the first region 14 and the second region 16 is a channel region. The channel region comprises two portions: a first portion 4 which is adjacent to the first region 14 and a second portion 6 which is adjacent to the second region 16. Spaced apart from the substrate 12 is a floating gate 18 which is on a insulating layer 20. The floating gate 18 is positioned over the second portion 6 of the channel region and is capacitively coupled to the second region 16. The floating gate 18 has a tip 22. A source contact 28 contacts the second region 16 while a drain contact 30 contacts the first region 14. Spaced apart and insulated from the floating gate 18 is a word line 24 or control gate 24. The word line 24/control gate 24 is also on the insulating layer 20 and is spaced apart from the substrate 12. The control gate 24 is positioned over the first portion 4 of the channel region and is closest to the tip 22. Thus far, the structure shown and described is fully disclosed in U.S. Patent No. 5,029,130.

[0012] In the improvement, the nonvolatile memory cell 10 further comprises an erase gate 26. The erase gate is spaced apart and is insulated from the word line 24. In addition, the erase gate 26 is insulated and spaced apart from the tip 22 of the floating gate 20. Between the erase gate 26 and the tip 22 of the floating gate 20 is an insulating material. The insulating material permits Fowler-Nordheim tunneling of electrons from the floating gate 18 to the erase gate 26. Because the erase gate 26 can be separately controlled, the functions of erase and read/program can be separated, thereby creating a greater degree of control over the program and read operations than over the erase operations.

[0013] Referring to Figure 1B, there is shown an array 210 of the memory cells 10 of the present invention. As shown in Figure 1B, the array 210 comprises a plurality of memory cells 10 arranged in a plurality of rows and columns. As is well known in the art, the cells 10 are arranged such that each pair of immediately adjacent cells are mirror images of one another. Thus, a first region 14 is common to a pair of cells 10A and 10B, whereas a second region 16 is common to a pair of cells 10B and 10C. Cells 10 that are arranged in the same row such as cells 10A, 10B and 10C have their first regions 14 connected in common. Cells 10 that are arranged in the same column, such as cells 10A and 10E, have their erase gates 26 connected in common, and have their second regions 16 connected in common. As is well known to those skilled in the art, the terms rows and columns can be interchanged.

[0014] In the operation of the memory array 210, the following voltages are applied.

[0015]

	Erase gate 26	Control Gate 24		First Region 14		Second Region 16
		Sel	Unused	Sel	Unused	
Erase	+12v	0	0	0	0	0
Program	1.4v	1.4v	0	.5v	1.8v	+12v
Read	1.6-2.2v	1.6-2.2v	0	.8v	0	0

[0016] Referring to Figure 3A, there is shown one step in the manufacturing of the nonvolatile memory cell 10 and array 210 of the present invention. The initial steps for creating the isolation regions are described in U.S. Patent No. 6,329,685 whose disclosure is incorporated herein by reference in its entirety. In particular, initially, the steps shown in Figures 1A-1 through 1C-4 and from 2A-4 through 2F-4 of U.S. Patent 6,329,685 are initially formed. As a result, as shown in Figure 3A, the second region 16 is formed with source contacts 28 made thereto. In addition, the floating gates 18 are formed, each having a sharp tip 22. Finally, an oxide layer 20 of approximately 30-50 angstrom is deposited. A layer of silicon nitride 32 is formed laterally to the floating gate 18. Although silicon nitride 32 is formed, silicon oxide may also be used.

[0017] Polysilicon 24 is then deposited. After the polysilicon 24 is deposited, through CMP or chemical mechanical polishing, it is etched to a level 34. Thereafter, the polysilicon 24 is anisotropically etched to level 36. The end point for the etch stop 36 is reached when on the periphery of the array (not shown) the etch stops on the STI in the periphery. The thickness of the remaining polysilicon 24 should be approximately 500 angstroms. The STI level in the periphery should be approximately 500 angstroms above silicon surface to serve as the end point. The STI level in the cell array area should be flat with the silicon surface. From the structure shown in Figure 3B, the silicon nitride 32 is removed by isotropic dry etch. However, silicon nitride 32 remains between the polysilicon 24 and the floating gate 18. The resultant structure is shown in Figure 3C.

[0018] The structure in Figure 3C is then subjected to CMP and etch back with a final thickness of the word line 24 being approximately 500 angstroms. The resultant structure is shown in Figure 3D.

[0019] The structure shown in Figure 3D is then dipped in hydrofluoric acid to remove a portion of the oxide spacer 36 to expose the tip 22. The structure is then subject to an IPO (interpoly oxide) deposition forming a tunnel oxide layer 38 on the polysilicon 24 and covering the tip 22. Alternatively, the hydrofluoric acid dip occurs immediately prior to the deposition of layer 32. A tunneling dielectric 38 is formed by a combination of thermal oxidation and CVD deposition. This dielectric 38 will eventually insulate the erase gate 26 from the floating gate 18 and the word line 24. The resultant structure is shown in Figure 3E.

[0020] Thereafter, polysilicon 26 is deposited everywhere in a conformal deposition. This polysilicon 26 will eventually form the erase gate. The polysilicon 26 deposited includes a region over the tunnel oxide 38 and over the word line 24 but is insulated therefrom. The resultant structure is shown in Figure 3F.

[0021] The polysilicon 26 is etched with an anisotropic etch. As is well known, the combination of conformal deposition followed by anisotropic etch results in a well defined spacer at the vertical edges 36. The polysilicon layer 26 is etched until the interpolyoxide layer 38 is reached. The resultant structure is shown in Figure 3G.

[0022] Using the erase gate 26 as a mask, the interpoly oxide 38 is then etched and the word line or control gate 24 poly is then etched and the word line oxide layer 20 is etched until the planar surface 8 of the substrate 12 is reached. The resultant structure is shown in Figure 3H.

[0023] Appropriate implantation and spacers are made using the edges of the erase agtes 26 and the word line gates 24 to form the LDD structure for the first region 14. The resultant structure is shown in Figure 3I.

[0024] Contact formation 30 is then made to the first region 14. The resultant structure is shown in Figure 3J.

[0025] As can be seen from Figure 1B, the erase gate 26 and the word line gate 24 in the same row are connected in the same direction. Thus, since the erase gate 26 is "over" the word line gate 24, there is difficulty in accessing the word gate 24 for strapping. Accordingly, periodically, in a preferred embodiment this is 128 cells, the word line gate 24 is extended in a direction such that strapping can occur. A top planar view of the array 210 is shown in Figure 4. As can be seen, where the cross-sectional line 3K-3K is shown, the erase gate 26 is "indented" permitting the word line 24 to be laterally extended to be connected to a strap. A cross-sectional view taken along the line 3K-3K is shown in Figure 3K showing the strapping to the word line 24.

[0026] Referring to Figure 2A, there is shown a second embodiment of a nonvolatile memory cell 110 of the present invention. The cell 110 is similar to the cell 10 shown in Figure 1A, with the exception that each cell 110 has two floating gates and operates bidirectionally. In addition, the cell 110 has a control gate 24 and a separate erase gate 26 for each cell 110. More specifically, each cell 110 has a first region 14 and a second region 16 spaced apart from one another with a channel region therebetween, in a substrate 12. Each of the first region 14 and a second region 16, however, lies in a trench in the substrate 12 with a portion of the channel region being a planar surface 8. Each of the trenches has a bottom and a side wall with the first region 14 and the second region 16 being at the bottom of the trench. A first floating gate 18A and a second floating gate 18B are along the side of the sidewalls, spaced apart therefrom and insulated therefrom. Thus, each of the floating gate 18A and 18B controls the conduction of the channel region which is along the sidewall of the trench. The channel region of the cell 110 is similar to the channel region of the cell 10 in that it has a first portion 4 and a second portion 6. The floating gate controls the second portion 6 of the channel which is along the sidewall of the trench. A control gate 24 is substantially parallel to the planar surface 8 and controls the first portion 4 of the channel region which is along the planar surface 8. A first contact 30 contacts the first region 14 and a second contact 28 contacts the second region 16. Thus, each of the contacts 30 and 28 extend into the trench. A contact 29 electrically connects to the control gate 24. Each of the floating gates 18A and 18B has a tip 22A and 22B respectively which are pointed away from the bottom of the trenches where the first region 14 and 16 lie. Thus, the tips 22A and 22B are adjacent to but spaced apart from the control gate 24. An erase gate 26 is substantially above the control gate 24 and is positioned to intercept electrons or charges emitted from the floating gate 18A and 18B. The erase gate 26 is insulated and separate from the control gate 24.

[0027] The operation of the cell 110 is as follows.

[0028] To program, the first region 14 is held at a small positive voltage such as 0.1 volts, the word line 24 is at a voltage sufficient to turn on the second portion 4 of the channel region, and the second region 16 is held at a programming voltage such as +6 volts. The erase gate 26 is held at a moderate positive voltage such as +3.0V. The voltages on the erase gate 26 and word lines 24 capacitively couple voltage on the first floating gate 18A. That voltage together with the initial charge state of the floating gate 18A are sufficient to invert the second portion 6 of the channel. With the second portion 6 of the channel region being turned on, and the first portion 4 being turned on, electrons are accelerated as they traverse to the second region 16 and are injected onto the second floating gate 18B through the mechanism of hot channel electron injection similar to the operation described for the cell 10 shown in Figure 1A. To program the first floating gate 18A, the voltages on the first region 14 and the second region 16 are reversed.

[0029] To erase both the first floating gate 18A and the second floating gate 18B, the first region 14 and the second region 16 are held at ground. The erase gate 26 is held at a high potential such as +12 volts. The control gate 24 is held at floating. In such a case, the electrons stored on the floating gate 18A and 18B are attracted by the high positive potential on the erase gate 26 and through the mechanism of Fowler-Nordheim tunneling, they tunnel through the interpoly oxide to the erase gate 26.

[0030] To read the cell 110 and to determine if the floating gate 18B is programmed, the second region 16 is held to ground. The word line 24 is held at +2 volts sufficient to turn on the first portion 4 of the channel region. A positive potential such as +3 volts is applied to the first region 14. With the first region 14 at +3 volts and the contact 30 at +3 volts, even if the first floating gate 18A were charged, the depletion region would extend to the first portion 4 of the channel region. Conduction of the electrons in the channel region between the first region 14 and the second region 16 would then depend upon the state of the floating gate 18B. If floating gate 18B were erased, then the channel region adjacent to the second floating gate 18B would conduct and a read current would pass from the first region 14 to the second region 16. If the second floating gate 18B were programmed, then the negatively charged electrons on the second floating gate 18B would prevent a read current from passing between the first region 14 and the second region 16. To read the cell 110 to determine whether the first floating gate 18A is programmed, the voltages applied to the first region 14 and the second region 16 are reversed. During read, a moderate voltage, +3.0V may be applied to the erase gate 26 to capacitively couple voltage to the floating gates 18A and 18B. This shifts the voltage operating window as may be convenient for circuit operation.

[0031] From the foregoing, it is seen that the operation of the cell 110 is similar to the operation of cell 10, except cell 110 operates bidirectionally.

[0032] A schematic view of an array 310 employing the cells 110 of the present invention is shown in Figure 2B. As shown in Figure 2B, the array 310 comprises a plurality of cells 110 arranged in a plurality of rows and columns. Again, the term "row" and "column" are interchangeable. For the cells 110 that are in the same row such as cells 110A, 110B and 110C, the erase gate 26 is connected together. In addition, the control gate 24 connects cells 110 that are in the same row. For cells that are in the same column, i.e. cells 110A, 110E and 110I, the contact line 30 connects the first regions 14 together. In addition, the contact line 28 connects the second region 16 together.

[0033] Consistent with the foregoing, the voltages applied to the various selected and unselected cells and portions thereof for the operations of program, erase and read are as follows:

[0034]

	Erase gate 26	Control Gate 24		First Region 14		Second Region 16	
		Sel	Unused	Sel	Unused	Sel	Unused
Erase	+12v	0	0	0	0	0	0
Program gate 18a	3-5v	V_t	0	+6v	0	0~0.5	0
Program gate 18b	3-5v	V_t	0	0~0.5	0	+6v	0
Read gate 18a	3-5v	2~4v	0	0	0	2~3v	0
Read gate 18b	3-5v	2~4v	0	2~3v	0	0	0

[0035] A method of manufacturing the cell 110 and the array 310 is as follows. Referring to Figure 5A there is shown a cross-sectional view of the first step in the process of making the cell 110 and the array 310. The layers of materials that are on the substrate 12 (having a first conductivity, typically P type) consist of an oxide layer 20 of approximately 20-50 angstroms, a polysilicon layer 24 of approximately 500-1000 angstroms (which may also be polysilicide), and an oxide layer 50 of approximately 500-1000 angstroms on the polysilicon 24. Finally, a layer 52 of 200-400 angstroms of silicon nitride is on the second layer of oxide 50.

[0036] The structure shown in Figure 5A is then subject to a masking operation in which portions of the structure are masked and the unmasked portions are etched. The etching occurs through the silicon nitride layer 52, the second layer of oxide 50, and the polysilicon 24. The resultant structure is shown in Figure 5B.

[0037] The photo resist (not shown) is then stripped from the region above the silicon nitride 52. Using the silicon nitride 52, the second oxide 50 and the polysilicon 24 as a mask, the first oxide region 20 is then etched and the underlying silicon substrate 12 is also etched to form trenches that are approximately 300-500 angstroms. The resultant structure is shown in Figure 5C.

[0038] The structure shown in Figure 5C is then subject to a shallow N+ diffusion implant forming the first region 14 and the second region 16. The resultant structure is shown in Figure 6D. Although the first and second regions 14 and 16 are shown as being implanted into the substrate 12, they can also be implanted into a well within a substrate 12. The implant can occur with Arsenic ions at $1 \times 10^{15}/cm^2$ dosage at 20keV. Thus, each of the first region 14 and the second region 16 forms a continuous buried diffusion line. The resultant structure is shown in Figure 5D.

[0039] A layer of oxide 54 is then conformally deposited onto the structure shown in Figure 5D. This would cover the substrate 12 as well as the "side" of the polysilicon 24. Alternatively, the "exposed" polysilicon 24 and silicon substrate 12 can be oxidized in situ to form silicon dioxide. A second layer of polysilicon 18 is then conformally deposited on the layer 54 of oxide. The resultant structure is shown in Figure 5E.

[0040] The polysilicon 18 is then anisotropically etched stopping at the oxide layer 54 forming the resultant first floating gate 18A and the second floating gate 18B. Each floating gate 18 has a tip 22. The oxide layer 54 serves to insulate the floating gate 18A and 18B from the control gate 24. The resultant structure is shown in Figure 5F.

[0041] The structure shown in Figure 5F is then subject to a high temperature oxide deposition step in which a layer 56 of high temperature deposit oxide is conformally deposited on the structure shown in Figure 5F. The structure is then subject to a CMP step stopping on the silicon nitride layer 52. The resultant structure is shown in Figure 5G.

[0042] A masking step using photoresist is then applied. Photoresist is applied across the surface of the structure shown in Figure 5G. A mask is applied exposing strips of photoresist that lie above and below the plane of the paper of Figure 5G. The strips of photoresist (either exposed or unexposed) are removed. The exposed nitride 52, oxide 54 are anisotropically removed. The exposed polysilicon 24 and polysilicon 18 are then removed thereby cutting their continuity. This forms rows of insulation, i.e., rows in which there is no polysilicon 24 and polysilicon 18, although the diffusion regions 14/16 continue to pass through the insulation row.

[0043] Middle Of Line oxide layer (MOL) 58, such as BPSG, is then deposited everywhere to a depth of approximately 500-1000 angstroms. The resulting structure is shown in Figure 5H.

[0044] CMP is then applied to the structure shown in Figure 5H with the silicon nitride layer 52 used as the polished stop. The resultant structure is shown in Figure 5I.

[0045] The silicon nitride 52 is then removed by wet etch or dry etch. Further, a short amount of an oxide wet etch is employed on the structure to further expose the tip 22A and 22B of the first and second floating gates 18A and 18B, respectively. The resultant structure is shown in Figure 5J.

[0046] A layer 60 of high temperature oxide is then deposited everywhere covering the tips 22A and 22B. The high temperature oxide is deposited to a depth of approximately 120-200 angstroms. Thereafter, polysilicon 26 to form the erase gate is deposited on the tunneling oxide 60. The resultant structure is shown in Figure 5K. The polysilicon 26 is a continuous sheet covering over many rows of cells, since the polysilicon 26 is the erase gate and a sector of cells can be erased at a time in a flash device.

[0047] Contacts are then formed to the structure shown in Figure 5K forming the cross-section view shown in Figure 2A. As is apparent, although three contacts per cell are formed, adjacent cells share the same either first region contact 30 or the second contact region 28. The contacts 30 and 28 contact the buried diffusion 14 and 16, respectively, through the polysilicon 26. Thus, they are used for strapping purpose and need not be applied to every row of cells. The metal lines (not shown) to which the contacts 30 and 28 are attached would run in a direction perpendicular to the paper as shown in Figure 2B. The contact 29 is also made through the polysilicon 26 and the polysilicon 24. They are made to each cell. The metal line (not shown) to which the contacts 29 are attached would run in a direction parallel to the row direction, as shown in Figure 2B. Because all the contacts 28, 29 and 30 are made through the polysilicon 26, the polysilicon 26 is fairly "holey." The structural integrity and electrically continuity of the polysilicon 26 is maintained by the polysilicon 26 being continuous over the adjacent insulation rows.